Semantic Role Labelling of Prepositional Phrases

Patrick Ye
Department of Computer Science and Software Engineering
University of Melbourne, VIC 3010, Australia

and

Timothy Baldwin
Department of Computer Science and Software Engineering
NICTA Victoria Research Laboratories
University of Melbourne, VIC 3010, Australia

---

1. INTRODUCTION

Prepositional phrases (PPs) are both common and semantically varied in open English text. Learning the semantics of prepositions is not a trivial task in general. It may seem that the semantics of a given PP can be predicted with reasonable reliability independent of its context. However, it is actually common for prepositions or even identical PPs to exhibit a wide range of semantic fuctions in different contexts. For example, consider the PP *to the car*: this PP will generally occur as a directional adjunct (e.g. *walk to the car*), but it can also occur as an object to the verb (e.g. *refer to the car*) or contrastive argument (e.g. *the default mode of transport has shifted from the train to the car*); to further complicate the situation, in *key to the car* it functions as a complement to the N-bar *key*. Based on this observation, we may consider the possibility of constructing a semantic tagger specifically for PPs, which uses the surrounding context of the PP to arrive at a semantic analysis. It is this task of PP semantic role labelling that we target in this paper.

A PP semantic role labeller would allow us to take a document and identify all adjunct PPs with their semantics. We would expect this to include a large portion of locative and temporal expressions, e.g., in the document, providing valuable data for tasks such as information extraction and question answering. Indeed our initial foray into PP semantic role labelling relates to an interest in geospatial and temporal analysis, and the realisation of the importance of PPs in identifying and classifying spatial and temporal references.

The contributions of this paper are to propose a method for PP semantic role labelling, and evaluate its performance over both the Penn Treebank (including comparative evaluation with previous work) and also the data from the CoNLL Semantic Role Labelling shared task. As part of this process, we identify the level of complementarity of a dedicated PP semantic role labeller with a conventional holistic semantic role labeller, suggesting PP semantic role labelling as a potential

Ye, Patrick and Timothy Baldwin (2006) Semantic Role Labelling of Prepositional Phrases, ACM Transactions on Asian Language Information Processing (TALIP), 5(3), pp. 228–244.

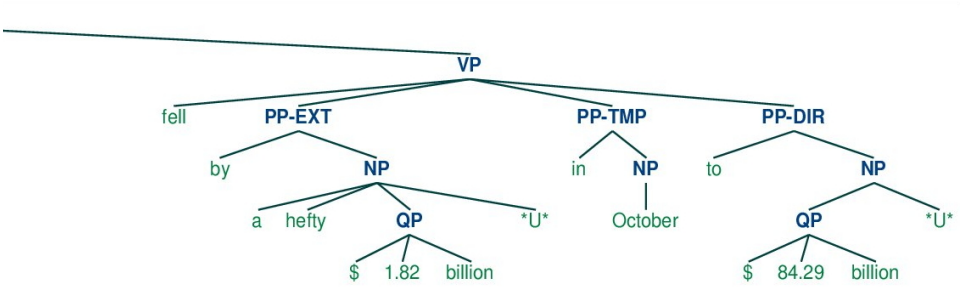

Fig. 1. An example of the preposition semantic roles in Penn Teebank

avenue for boosting the performance of existing systems.

In the remainder of this paper, we outline the propose a method for PP semantic role disambiguation, and evaluate it over both the Penn Treebank (Section 2) and the CoNLL 2004 Semantic Role Labelling shared task (Section 3). We then contrast the relative success of the proposed method over the two data sets (Section 4), and finally conclude the paper with a discussion of future work (Section 5).

2. PREPOSITION SEMANTIC ROLE DISAMBIGUATION IN PENN TREEBANK

Significant numbers of prepositional phrases (PPs) in the Penn Treebank [Marcus et al. 1993] are tagged with their semantic role relative to the governing verb. For example, Figure 1 shows a fragment of the parse tree for the sentence *[Japan's reserves of gold, convertible foreign currencies, and special drawing rights] fell by a hefty $1.82 billion in October to $84.29 billion [the Finance Ministry said]*, in which the three PPs governed by the verb *fell* are tagged as, respectively: `PP-EXT` ("extend"), meaning how much of the reserve fell; `PP-TMP` ("temporal"), meaning when the reserve fell; and `PP-DIR` ("direction"), meaning the direction of the fall.

According to our analysis, there are 143 preposition semantic roles in the treebank. However, many of these semantic roles are very similar to one another; for example, the following semantic roles were found in the treebank: `PP-LOC`, `PP-LOC-1`, `PP-LOC-2`, `PP-LOC-3`, `PP-LOC-4`, `PP-LOC-5`, `PP-LOC-CLR`, `PP-LOC-CLR-2`, `PP-LOC-CLR-TPC-1`. Inspection of the data revealed no systematic semantic differences between these PP types. Indeed, for most PPs, it was impossible to distinguish the subtypes of a given superclass (e.g. `PP-LOC` in our example). We therefore decided to collapse the PP semantic roles based on their first semantic feature. For example, all semantic roles that start with `PP-LOC` are collapsed to the single class `PP-LOC`. Table I shows the distribution of the collapsed preposition semantic roles.

2.1 System Description

O'Hara and Wiebe [2003] describe a system[1] for disambiguating the semantic roles of prepositions in the Penn Treebank according to 7 basic semantic classes. In their

[1] This system was trained with WEKA's J48 decision tree implementation.

Semantic Role	Count	Frequency	Meaning
PP-LOC	21106	38.2	Locative
PP-TMP	12561	22.7	Temporal
PP-CLR	11729	21.2	"Closely related" (somewhere between an argument and an adjunct)
PP-DIR	3546	6.4	Direction (*from/to* X)
PP-MNR	1839	3.3	Manner (incl. instrumentals)
PP-PRD	1819	3.3	Predicate (non-VP)
PP-PRP	1182	2.1	Purpose or reason
PP-CD	654	1.2	Cardinal (numeric adjunct)
PP-PUT	296	0.5	Locative complement of *put*

Table I. Penn Treebank semantic role distribution (top-9 roles)

system, O'Hara and Wiebe used a decision tree classifier, and the following types of features:

—**POS tags of surrounding tokens:** The POS tags of the tokens before and after the target preposition within a predefined window size. In O'Hara and Wiebe's work, this window size is 2.

—**POS tag of the target preposition**

—**The target preposition**

—**Word collocation:** All the words in the same sentence as the target preposition; each word is treated as a binary feature.

—**Hypernym collocation:** The WordNet hypernyms [Miller 1995] of the open class words before and after the target preposition within a predefined window size (set to 5 words); each hypernym is treated as a binary feature.

O'Hara and Wiebe's system also performs the following pre-classification filtering on the collocation features:

—**Frequency constraint:** $f(coll) > 1$, where *coll* is either a word from the word collocation or a hypernym from the hypernym collocation

—**Conditional independence threshold:** $\frac{p(c|coll)-p(c)}{p(c)} >= 0.2$, where c is a particular semantic role and *coll* is from the word collocation or a hypernym from the hypernym collocation

We began our research by replicating O'Hara and Wiebe's method and seeking ways to improve it. Our initial investigation revealed that there were around 44000 word and hypernym collocation features even after the frequency constraint filter and the conditional independence filter have been applied. We did not believe all these collocation features were necessary, and deployed an additional frequency-threshold-based filtering mechanism over the collocation features to only select collocation features which occur in the top N frequency bins.

This frequency-threshold-based filtering mechanism allows us to select collocation feature sets of differing size, and in doing so not only improve the training and tagging speed of the preposition semantic role labelling, but also observe how the number of collocation features affects the performance of the PP semantic role labeller and which collocation features are more important.

Top N most Frequent Features	Accuracy (%) Classifier 1	Classifier 2
10	74.75	81.28
20	76.53	83.52
50	79.21	86.34
100	80.13	87.02
300	81.32	87.62
1000	82.34	87.71
all	82.76	87.45
O'Hara & Wiebe	N/A	85.8

Table II. Penn Treebank preposition semantic role disambiguation results

2.2 Results

Since some of the preposition semantic roles in the treebank have extremely low frequencies, we decided to build our first classifier using only the top 9 semantic roles, as detailed in Table I. We also noticed that the semantic roles `PP-CLR, PP-CD and PP-PUT` were excluded from O'Hara's system which only used `PP-BNF, PP-EXT, PP-MNR, PP-TMP, PP-DIR, PP-LOC and PP-PRP`, and therefore built a second classifier using only the semantic roles used by O'Hara's system.[2] The two classifiers were built with a maximum entropy [Berger et al. 1996] learner.[3]

Table II shows the results of our classifier under stratified 10-fold cross validation[4] using different parameters for the ranking-based filter. We also list the accuracy reported by O'Hara and Wiebe for comparison.

The results show that the performance of the classifier increases as we add more collocation features. However, this increase is not linear, and the improvement of performance is only marginal when the number of collocation features is greater than 100. It also can be observed that there is a consistent performance difference between classifiers 1 and 2, which suggests that `PP-CLR` may be harder to distinguish from the other semantic roles. This is not totally surprising given the relatively vague definition of the semantics of `PP-CLR`. We return to analyze these results in greater depth in Section 4.

3. PREPOSITION SEMANTIC ROLE LABELLING OVER THE CONLL 2004 DATA SET

Having built a classifier which has reasonable performance on the task of treebank preposition semantic role disambiguation, we decided to investigate whether we could use a similar set of features to perform PP semantic role labelling over alternate systems of PP classification. We chose the 2004 CoNLL Semantic Role Labelling (SRL) data set [Carreras and Màrquez 2004] because it contained a wide range of semantic classes of PPs, in part analogous to the Penn Treebank data, and also because we wished to couple our method with a holistic SRL system to demonstrate the ability of PP semantic role labelling to enhance overall system performance.

[2] `PP-BNF` with only 47 counts was not used by the second classifier.

[3] `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`

[4] O'Hara's system was also evaluated using stratified 10-fold cross validation.

Since the focus of the CoNLL data is on SRL relative to a set of pre-determined verbs for each sentence input,[5] our primary objective is to investigate whether the performance of SRL systems in general can be improved in any way by an independent preposition SRL system. We achieve this by embedding our PP classification method within an existing holistic SRL system—that is a system which attempts to tag all semantic role types in the CoNLL 2004 data—through the following three steps:

(1) Perform SRL on each preposition in the CoNLL data set;
(2) Merge the output of the preposition SRL with the output of a given verb SRL system over the same data set;
(3) Perform standard CoNLL SRL evaluation over the merged output.

The details of preposition SRL and combination with the output of a holistic SRL system are discussed below.

3.1 Breakdown of the Preposition Semantic Role Labelling Problem

Preposition semantic role labelling over the CoNLL data set is considerably more complicated than the task of disambiguating preposition semantic roles in the Penn Treebank. There are three separate subtasks which are required to perform preposition SRL:

(1) **PP Attachment:** determining which verb to attach each preposition to.
(2) **Preposition Semantic Role Disambiguation**
(3) **Argument Segmentation:** determining the boundaries of the semantic roles.

The three subtasks are not totally independent of each other, as we demonstrate in Section 3.8, and improved performance over one of the subtasks does not necessarily correlate with an improvement in the final results.

3.2 Preposition Verb Attachment Classification

Preposition-Verb attachment (VA) classification is the first step of preposition semantic role labelling and involves determining the verb attachment site for a given preposition, i.e. which of the pre-identified verbs in the sentence the preposition is governed by.

Verb Attachment Classification Using a Maximum Entropy Classifier

This classifier uses the following features, all of which are derived from information provided in the CoNLL data:

—**POS tags of surrounding tokens:** The POS tags of the tokens before and after the target preposition within a window size of 2 tokens ($[-2, 2]$).
—**POS tag of the target preposition**
—**The target preposition**

[5] Note that the CoNLL 2004 data identifies certain verbs as having argument structure, and that the semantic role annotation is relative to these verbs only. This is often not the sum total of all verbs in a given sentence: the verbs in relative clauses, e.g., tend not to be identified as having argument structure.

VA	Count	Frequency
None	3005	60.71
-1	1454	29.37
1	411	8.30
-2	40	0.81
2	29	0.59
3	8	0.16
-3	2	0.04
-6	1	0.02

Table III. VA class distribution

—**Verbs and their relative position (VerbRelPos):** All the (pre-identified) verbs in the same sentence as the target preposition and their relative positions to the preposition are extracted as features. Each (verb, relative position) tuple is treated as a binary feature. The relative positions are determined in a way such that the 1st verb before the preposition will be given the position -1, the 2nd verb before the preposition will be given the position -2, and so on.
—**The type of the clause containing the target preposition**
—**Neighbouring chunk type:** The types (NP, PP, VP, etc.) of chunks before and after the target preposition within a window of 3 chunks.
—**Word collocation (WordColl):** All the open class words in the phrases before and after the target preposition within a predefined window of 3 chunks.
—**Hypernym collocation (HyperColl):** All the WordNet hypernyms from all the senses of the open class words in the phrases before and after the target preposition within a predefined window of 3 chunks.
—**Named Entity collocation NEColl:** All the named entity information from the phrases before and after the target preposition within a predefined window of 3 chunks.
—**Chunk-based N-Gram features:** A series of N-gram features were used to capture the more abstract syntactic and contextual features around the relevant preposition. In this study, the first 5 chunks after the relevant preposition were used to derive these features. These features are:
  —**Regular expression representation of the chunk types:** This feature is created by merging consecutive identical chunk types into a single symbol. For example, the chunk sequence **VP NP PP NP NP** will be represented as **VP_NP_PP_NP+**.
  —**The first word of each chunk**
  —**The last word of each chunk**
  —**The first part of speech tag of each chunk**
  —**The last part of speech tag of each chunk**

The VA classifier outputs the relative position of the governing verb to the target preposition, or `None` if the preposition does not have a semantic role. Such prepositions include those which are attached to noun phrases and those which are attached to verb phrases but are not semantically labelled by CoNLL 2004.

We trained the VA classifier over the CoNLL 2004 training set, and tested it on the testing set. Table III shows the distribution of the classes in the testing set.

Algorithm 1 Verb Attachment Using Charniak Parser

(1) Let w be the preposition for which we wish to find the parent
(2) Let p_w be the position of w in the parse tree
(3) Let p_{pp} be the parse tree position of the **prepositional phrase** that w is a direct child of
(4) Let p_x be the direct parent of p_{pp}
(5) Repeat the following
  (a) if p_x is a VP, then break
  (b) else if p_x is an NP or SBAR, then terminate and return None as w's verb attachment
  (c) else re-assign p_x to be its direct parent
(6) Let C be the set of p_x's terminal children ordered according to their positions in the original sentence from left to right, then do the following:
  (a) let $v = None$
  (b) for c in C
    i. if c is w or to the right of w and $v \neq None$, then terminate and return v as the verb w is attached to
    ii. else if c is a verb then $v = c$
(7) Return $None$ as w's verb attachment

The same maximum entropy learner used in the treebank SRL task was used to train the VA classifier. The accuracy of this classifier on the CoNLL 2004 testing set is 80.14%.

Verb Attachment Classification Using the Charniak Parser

We also experimented with the Charniak parser [Charniak 2000] in the verb-preposition attachment classification. Since this parser was trained on the Penn Treebank data, which was also the source for the CoNLL 2004 data, we expect its accuracy to be reasonably high.

In this experiment, the parser was used to identify which verb a given preposition was attached to, or whether the given preposition was attached to a verb at all. However, it must be noted that since the parse trees produced by the Charniak parser do not contain any semantic role information, it would not be possible to distinguish prepositions which have semantic roles from prepositions that do not have semantic roles. Algorithm 1 shows the process of the verb-preposition attachment extraction.

Using the Charniak parser, the accuracy of the verb-preposition attachment classification is 71.19%.

Verb Attachment Classification Error Analysis

Table III shows that more than half of the prepositions classified by the verb attachment classifier actually did not have any semantic roles. In other words, most of the prepositions in the VA classification will not play a direct role in determining the performance of the entire preposition SRL system. Therefore, these prepositions are not as important as the ones that do have semantic roles. However, this

VA	Count	Frequency %
-1	1454	74.76
1	411	21.13
-2	40	2.06
2	29	1.49
3	8	0.41
-3	2	0.10
-6	1	0.05

Table IV. Revised VA class distribution in test set

VA	Maxent Classifier Accuracy %	Parse Tree Classifier Accuracy %
None	88.65	71.19
-1	73.87	78.95
1	55.47	0.97
-2	2.50	60.00
2	0.00	0.00
-3	0.00	0.00
3	0.00	0.00
-6	0.00	0.00
overall with *None*	80.14	78.13
overall without *None*	66.99	60.46

Table V. Breakdown of the accuracy of the VA classifiers on the test set

factor was not taken into account by the naive accuracy metric used to measure the performance of the VA classifier, and as a result, the naive accuracy metric may not be able to accurately reflect the real difficulty of the VA classification task.

To address this issue, we re-analyzed the performance of the VA classifier by only looking at its accuracy on prepositions which have semantic roles. Table IV shows the distribution of verb-preposition attachments without the *None* classification. Table V shows the breakdown of the verb-preposition attachment classification accuracy on the test data set of both the maxent based classifier and the Charniak parser based classifier.

One interesting observation that can be made from Table V is that the parse tree based classifier performed extremely poorly when the preposition was attached to the first verb **after** it. This suggests that either the parser did a poor job on these sentences, or there is a flaw in Algorithm 1.

Recall that the *None* classification is assigned to prepositions not attached to any verbs, which therefore have no direct impact on the preposition SRL task as a whole. However, since these prepositions account for the majority of the data, if they are included in the classification, the accuracy would be appear to be higher. Hence, we are much more interested in the classification accuracy of the prepositions which are actually attached to verbs. As Table V shows, the classification accuracy of these prepositions is rather poor, and as a result, in the best case scenario, the overall preposition SRL system can only achieve an accuracy of 66.99%. It would be highly desirable to significantly improve this upper bound.

Another interesting observation about Table V is that the two VA classifiers performed quite differently with respect to the different verb-preposition attachments.

VA	Accuracy %
None	92.18
-1	87.14
1	54.26
-2	42.50
2	0.00
-3	0.00
3	0.00
-6	0.00
overall with *None*	86.40
overall without *None*	77.48

Table VI. Breakdown of the accuracy of the new maxent VA classifiers on the test set

This means that there is a certain level of difference in the capabilities of these two classifiers. Therefore, even though the parse tree classifier performs noticeably poorer than the maxent classifier, it is quite possible that a significant portion of the mistakes made by the two classifiers are actually made on different test instances. A further analysis of the mistakes made by the two classifiers confirmed this: for the test set, if the accuracy was calculated in a way such that an example is considered correctly classified when **one of the two** classifiers produces the right classification, then the overall accuracies with and without the *None* classification would respectively become 92.16% and 83.24%. This would be a much better upper bound for the accuracy of the overall preposition SRL system.

A New Maxent Classifier Incorporating the Parse Tree Classifier for Verb Attachment

In order to take advantage of the different strengths of the two existing VA classifiers, we constructed a new maxent classifier using all the features of the first maxent classifier, and one additional feature: the **classification result of the parse tree VA classifier**. Table VI shows the breakdown of the accuracies of this new maxent classifier, and it is obvious that this new maxent classifier performs much better than both the old maxent classifier and the parse tree classifier, and it was therefore used in the final preposition SRL system.

3.3 Preposition Semantic Role Disambiguation

For the task of preposition semantic role disambiguation (SRD), we constructed a classifier using the same features as the VA classifier, with the following differences and additional features:

(1) The window size for the POS tags of surrounding tokens is 5 tokens.
(2) The window sizes for the **WordColl**, the **HyperColl** and the **NEColl** features are set to include the entire sentence.
(3) The chunk tags (in the IOB format [Tjong Kim Sang and Veenstra 1999]) of the words within a window of 5.

We trained the SRD classifier once again on the CoNLL 2004 training set, and tested it on the testing set. Table VII shows the distribution of the classes in the testing set.

We used the same maximum entropy learner as for the VA classifier to train the SRD classifier. The accuracy of the SRD classifier on the CoNLL 2004 testing set

Semantic Role	Count	Frequency	Meaning
A1	424	21.79	Argument 1
A2	355	18.24	Argument 2
AM-TMP	299	15.36	Temporal adjunct
AM-LOC	188	9.66	Locative adjunct
A0	183	9.40	Argument 0
AM-MNR	125	6.42	Manner adjunct
A3	106	5.45	Argument 3
AM-ADV	71	3.65	General-purpose adjunct
A4	44	2.26	Argument 4
AM-CAU	40	2.06	Causal adjunct
AM-PNC	32	1.64	Purpose adjunct
AM-DIS	32	1.64	Discourse marker
AM-DIR	19	0.97	Directional adjunct
AM-EXT	7	0.36	Extent adjunct

Table VII. CoNLL 2004 semantic role distribution in the CoNLL 2004 test data set(top-14 roles)

Algorithm 2 Regular Expression Based Segmentation Algorithm

(1) Let s be the index of the start of the preposition chunk
(2) Let e be the index of the end of the preposition chunk
(3) Go through the chunks following the preposition chunk and assign their end index to be e until one of the following conditions is satisfied:
  (a) The end of the sentence is reached
  (b) A preposition which is attached to a verb is reached
  (c) A chunk which is not an NP chunk is reached.

is 63.36%.

3.4 Argument Segmentation

Once the semantic role and verb attachment of a preposition has been determined, it would then be necessary to determine the boundary of the semantic role, i.e. argument segmentation. For this task, we have experimented with both a simple regular expression based method and a more complex statistical parser approach. The details are given below.

Argument Segmentation Using A Regular Expression

This method determines the extent of each NP selected for by a given preposition (i.e. the span of words contained in the NP), and is based on a simple regular expression (RE) over the chunk parser analysis of the sentence provided in the CoNLL 2004 data, namely: `PP` $\texttt{NP}^{+}$. The details of this algorithm are shown in Algorithm 2.

The performance of the regular expression based argument segmentation cannot be independently evaluated. This is because the segmentation convention used by the CoNLL 2004 data seems to differentiate between **argument type** semantic roles (such as $A0$, $A1$) and **modifier type** semantic roles (such as $AM-LOC$, $AM-TMP$). In the case of an **argument type**, the semantic role boundary will start from the preposition, but in the case of a **modifier type**, the semantic role

Algorithm 3 Charniak Parser Based Segmentation Algorithm

(1) Let w be the preposition of interest
(2) Let t be a sibling of w in the parser tree, where t is immediately to the right of w
(3) If t is a nonterminal, then the boundary of the argument starts from the first terminal child of t and ends at the last terminal child of t
(4) Else, the boundary of the argument starts from and ends at t

boundary will start from the first word after the preposition. During the boundary extraction process, the segmentation module has no access to the semantic role information, so it is not possible to determine where exactly the argument boundary should start, and therefore the first word after the preposition of interest is always assigned to be the start of the argument boundary. In the process of combining the final outputs of all the subtasks, we then use the semantic role information produced by the preposition SRD module to finally determine where exactly the relevant argument should start.

Based on the above, for the purpose of evaluation, we decided to use the perfect preposition SRD results to first compensate the output of the segmentation module, then compare it against the correct segmentation. The accuracy of the regular expression based segmentation method is 53.08%.

Argument Segmentation Using Statistical Parsers

We realized that the RE based segmentation method was only capable of extracting arguments which were just noun phrases, and was not robust enough as a result of this limitation. Therefore we decided to experiment with the Charniak parser and the RASP parser [Briscoe and Carroll 2002] to see if better segmentation results could be achieved.

Similar to the RE method, we assumed that the boundary of the argument starts from the first word after the preposition of interest. Algorithm 3 shows how the parse trees are used to perform the task of argument segmentation.

The evaluation of the parser based segmentation method was performed in the same way as the RE segmentation method. The Charniak parser based classifier achieved an accuracy of 71.48%, and the RASP based classifier achieved an accuracy of 50.05%. Since the Charniak parser based classifier worked significantly better than the other two methods, it was used in the final preposition SRL system.

We were not surprised by the significant gap between the performances of the Charniak parser and RASP. As stated before, the Charniak parser was trained over a superset of the CoNLL 2004 data, whereas RASP was trained on independent data.

3.5 Combining the Output of the Subtasks

Once we have identified the association between verbs and prepositions, and disambiguated the semantic roles of the prepositions, we can begin the process of creating the final output of the preposition semantic role labelling system. This takes place by identifying the data column corresponding to the verb governing each classified PP in the CoNLL data format (as determined by the VA classifier), and recording

	SRD_{AUTO}						SRD_{ORACLE}					
	SEG_{AUTO}			SEG_{ORACLE}			SEG_{AUTO}			SEG_{ORACLE}		
	P	R	F	P	R	F	P	R	F	P	R	F
VA_{AUTO}	48.71	7.65	13.22	62.80	10.33	17.73	74.50	11.69	20.21	<u>94.54</u>	15.51	26.65
VA_{ORACLE}	46.82	8.84	14.87	63.21	12.69	21.14	73.91	13.93	23.44	<u>99.38</u>	19.91	33.17

Table VIII. Preposition SRL results before merging with the holistic SRL systems, (P = precision, R = recall, F = F-score; above-baseline results <u>underlined</u>)

the semantic role of that PP (as determined by the SRD classifier) over the full extent of the PP (as determined by the segmentation classifier).

3.6 Parameter Tuning of the Maxent Based Classifiers

Since the maxent based machine learning package can be tuned based on the number of iterations i and the Gaussian prior smoothing parameter g, we decided to train all the maxent based classifiers on the training set of the CoNLL 2004 data with a wide range of combinations of the two parameters. We then applied each combination on the development set, then chose the best one to apply to the test set of the data.

3.7 Merging the Output of Preposition SRL and Verb SRL

Once we have generated the output of the preposition SRL system, we can proceed to the final stage where the semantic roles of the prepositions are merged with the semantic roles of an existing holistic SRL system.

It is possible, and indeed likely, that the semantic roles produced by the two systems will conflict in terms of overlap in the extent of labelled constituents and/or the semantic role labelling of constituents. To address any such conflicts, we designed three merging strategies to identify the right balance between the outputs of the two component systems:

S1. When a conflict is encountered, only use the semantic role information from the holistic SRL system.

S2. When a conflict is encountered, if the start positions of the semantic role are the same for both SRL systems, then replace the semantic role of the holistic SRL system with that of the preposition SRL system, but keep the holistic SRL system's boundary end.

S3. When a conflict is encountered, only use the semantic role information from the preposition SRL system.

3.8 Results

To evaluate the performance of our preposition SRL system, we combined its outputs with the 3 top-performing holistic SRL systems from the CoNLL 2004 SRL shared task.[6] The three systems are Hacioglu et al. [2004], Punyakanok et al. [2004] and Carreras et al. [2004]. Furthermore, in order to establish the upper bound of the improvement of preposition SRL on verb SRL, and investigate how the three

[6] Using the test data outputs of the three systems made available at `http://www.lsi.upc.edu/~srlconll/st04/st04.html`.

	SRD_{AUTO}						SRD_{ORACLE}					
	SEG_{AUTO}			SEG_{ORACLE}			SEG_{AUTO}			SEG_{ORACLE}		
	P	R	F	P	R	F	P	R	F	P	R	F
ORIG	72.43	66.77	69.49	72.43	66.77	69.49	72.43	66.77	69.49	72.43	66.77	69.49
S1												
VA_{AUTO}	72.20	66.98	69.50	72.12	67.01	69.48	72.36	67.13	69.65	72.41	67.27	69.75
VA_{ORACLE}	72.15	67.21	69.59	72.05	67.45	69.67	72.54	67.58	69.97	72.83	68.18	70.43
S2												
VA_{AUTO}	71.00	65.88	68.34	70.68	65.69	68.10	73.47	68.16	70.72	73.68	68.45	70.97
VA_{ORACLE}	70.68	65.85	68.18	70.17	65.70	67.86	73.95	68.87	71.32	74.43	69.66	71.97
S3												
VA_{AUTO}	70.96	64.85	67.77	73.43	68.23	70.74	75.42	68.90	72.01	79.07	73.43	76.15
VA_{ORACLE}	70.59	64.42	67.36	74.00	69.67	71.77	76.22	69.52	72.72	81.73	76.90	79.24

Table IX. Preposition SRL combined with Hacioglu et al. [2004] (P = precision, R = recall, F = F-score; above-baseline results underlined)

	SRD_{AUTO}						SRD_{ORACLE}					
	SEG_{AUTO}			SEG_{ORACLE}			SEG_{AUTO}			SEG_{ORACLE}		
	P	R	F	P	R	F	P	R	F	P	R	F
ORIG	70.07	63.07	66.39	70.07	63.07	66.39	70.07	63.07	66.39	70.07	63.07	66.39
S1												
VA_{AUTO}	69.21	64.89	66.98	69.34	65.25	67.24	70.58	66.16	68.30	71.04	66.83	68.87
VA_{ORACLE}	68.93	65.21	67.02	69.16	65.93	67.51	70.84	66.99	68.86	71.70	68.33	69.97
S2												
VA_{AUTO}	68.50	64.24	66.30	68.49	64.46	66.41	71.95	67.45	69.63	72.57	68.27	70.36
VA_{ORACLE}	68.05	64.39	66.17	67.95	64.78	66.33	72.55	68.61	70.52	73.62	70.15	71.84
S3												
VA_{AUTO}	67.79	62.36	64.96	70.17	65.72	67.87	72.25	66.43	69.22	75.79	70.94	73.29
VA_{ORACLE}	67.01	61.74	64.27	70.49	67.08	68.74	72.61	66.87	69.62	78.17	74.33	76.20

Table X. Preposition SRL combined with Punyakanok et al. [2004] (P = precision, R = recall, F = F-score; above-baseline results underlined)

	SRD_{AUTO}						SRD_{ORACLE}					
	SEG_{AUTO}			SEG_{ORACLE}			SEG_{AUTO}			SEG_{ORACLE}		
	P	R	F	P	R	F	P	R	F	P	R	F
ORIG	71.81	61.11	66.03	71.81	61.11	66.03	71.81	61.11	66.03	71.81	61.11	66.03
S1												
VA_{AUTO}	70.89	62.90	66.66	70.85	63.17	66.79	72.41	64.23	68.08	72.74	64.85	68.57
VA_{ORACLE}	70.55	63.43	66.80	70.61	64.07	67.18	72.66	65.31	68.79	73.51	66.68	69.93
S2												
VA_{AUTO}	70.33	62.40	66.13	70.13	62.53	66.12	73.52	65.21	69.12	74.05	66.00	69.80
VA_{ORACLE}	69.73	62.69	66.02	69.46	63.02	66.08	73.94	66.44	69.99	74.99	68.01	71.33
S3												
VA_{AUTO}	69.79	60.84	65.00	72.38	64.34	68.12	74.47	64.90	69.36	78.29	69.55	73.66
VA_{ORACLE}	68.99	60.50	64.47	72.64	65.89	69.10	74.86	65.62	69.94	80.68	73.13	76.72

Table XI. Preposition SRL combined with Carreras and Màrquez [2004] (P = precision, R = recall, F = F-score; above-baseline results underlined)

subtasks interact with each other and what their respective limits are, we also used oracled outputs from each subtask in combining the final outputs of the preposition SRL system. The oracled outputs are what would be produced by perfect classifiers, and are emulated by inspection of the gold-standard annotations for the testing data.

Table VIII shows the results of the preposition SRL systems before they are merged with the verb SRL systems. These results show that the coverage of our preposition SRL system is quite low relative to the total number of arguments in the testing data, even when oracled outputs from all three subsystems are used (recall = 18.15%). However, this is not surprising because we expected the majority of semantic roles to be noun phrases.

In Tables IX, X and XI, we show how our preposition SRL system performs when merged with the top 3 systems under the 3 merging strategies introduced in Section 3.7. In each table, ORIG refers to the base system without preposition SRL merging.

We can make a few observations from the results of the merged systems. First, out of verb attachment, SRD and segmentation, the SRD module is both: (a) the component with the greatest impact on overall performance, and (b) the component with the greatest differential between the oracle performance and classifier (AUTO) performance. This would thus appear to be the area in which future efforts should be concentrated in order to boost the performance of holistic SRLs through preposition SRL.

Second, the results show that in most cases, the recall of the merged system is higher than that of the original SRL system. This is not surprising given that we are generally relabelling or adding information to the argument structure of each verb, although with the more aggressive merging strategies (namely S2 and S3) it sometimes happens that recall drops, by virtue of the extent of an argument being adversely affected by relabelling. It does seem to point to a complementarity between verb-driven SRL and preposition-specific SRL, however.

There are a few aberrations in the results. Sometimes, an all-auto method achieved better results than when one of the subtasks was oracled. For example, in Table IX, when merge scheme 1 was used, the all-auto combination yielded a precision of 72.20%, and an F-score of 69.50%, whereas when the segmentation was substituted to oracled results, the precision dropped to 72.12% and the F-score dropped to 69.48%. This behaviour is caused by the poor accuracy of the SRD classifier and the merging strategy. The perfect segmentation results would reduce the number of argument boundary conflicts between the preposition SRL system and the original system, thereby increasing the recall of the combined system. However, due to the poor performance of the preposition SRD subsystem, these additional arguments were not correctly classified, and this was why the precision dropped while the recall improved.

Finally, it was somewhat disappointing to see that in no instance did a fully-automated method significantly surpass the base system in precision or F-score. Having said this, we were encouraged by the size of the margin between the base systems and the fully oracle-based systems, as it supports our base hypothesis that preposition SRL has the potential to boost the performance of holistic SRL systems, up to a margin of 10% in F-score for S3.

4. ANALYSIS AND DISCUSSION

In the previous two sections, we presented the methodologies and results of two systems that perform statistical analysis on the semantics of prepositions, each using

a different data set. The performance of the two systems was very different. The SRD system trained on the treebank produced highly creditable results, whereas the SRL system trained on CoNLL 2004 SRL data set produced somewhat negative results. In the remainder of this section, we will analyze these results and discuss their significance.

There is a significant difference between the results obtained by the treebank classifier and that obtained by the CoNLL SRL classifier. In fact, even with a very small number of collocation features, the treebank classifier still outperformed the CoNLL SRL classifier. This suggests that the semantic tagging of prepositions is somewhat artificial. This is evident in three ways. First, the proportion of prepositional phrases tagged with semantic roles is small – around 57,000 PPs out of the million-word treebank corpus. This small proportion suggests that the preposition semantic roles were tagged only in certain prototypical situations. Second, we were able to achieve reasonably high results even when we used a collocation feature set with fewer than 200 features. This further suggests that the semantic roles were tagged for only a small number of verbs in relatively fixed situations. Third, the preposition SRD system for the CoNLL data set used a very similar feature set to the treebank system, but was not able to produce anywhere near comparable results. Since the CoNLL data set is aimed at holistic SRL across all argument types, it incorporates a much larger set of verbs and tagging scenarios; as a result, the semantic role labelling of PPs is far more heterogeneous and realistic than is the case in the treebank. Therefore, we conclude that the results of our treebank preposition SRD system are not very meaningful in terms of predicting the success of the method at identifying and semantically labelling PPs in open text.

A few interesting facts came out of the results over the CoNLL data set. The most important one is that by using an independent preposition SRL system, the results of a general verb SRL system can be significantly boosted. This is evident because when the oracled results of all three subtasks were used, the merged results were around 10% higher than those for the original systems, in all three cases. Unfortunately, it was also evident from the results that we were not successful in automating preposition SRL. Due to the strictness of the CoNLL evaluation, it was not always possible to achieve a better overall performance by improving just one of the three subsystems. For example, in some cases, worse results were achieved by using the oracled results for VA and the results produced by SRD classifier, than using the VA classifier and the SRD classifiers in conjunction. The reason for the worse results is that in our experiments, the oracled VA always identifies more prepositions attached to verbs than the VA classifier. Therefore more prepositions will be given semantic roles by the SRD classifier, thus increasing the recall of the final system. However, since the performance of the SRD classifier is not high, and the segmentation subsystem does not always produce the same semantic role boundaries as the CoNLL data set, most of these additional prepositions would either be given a wrong semantic role or wrong phrasal extent (or both), thereby causing the overall performance to fall.

Finally, it is evident that the merging strategy also plays an important role in determining the performance of the merged preposition SRL and verb SRL systems: when the performance of the preposition SRL system is high, a more preposition-

oriented merging scheme would produce better overall results, and vice versa.

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a method for labelling preposition semantics and deployed the method over two different data sets involving preposition semantics. We have shown that preposition semantics is not a trivial problem in general, and also that it has the potential to complement other semantic analysis tasks, such as semantic role labelling.

Our analysis of the results of the preposition SRL system shows that significant improvement in all three stages of preposition semantic role labelling—namely verb attachment, preposition semantic role disambiguation and argument segmentation—must be achieved before preposition SRL can make a significant contribution to holistic SRL. The unsatisfactory results of our CoNLL preposition SRL system show that the relatively simplistic feature sets used in our research are far from sufficient. Therefore, we will direct our future work towards using additional NLP tools, information repositories and feature engineering to improve all three stages of preposition semantic role labelling.

Acknowledgements

We would like to thank Phil Blunsom and Steven Bird for their suggestions and encouragement, Tom O'Hara for providing insight into the inner workings of his semantic role disambiguation system, and the anonymous reviewers for their comments. National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology, and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Research Centre of Excellence Programs.

REFERENCES

Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics 22,* 1, 39–71.

Briscoe, T. and Carroll, J. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Canary Islands, 1499–1504.

Carreras, X. and Màrquez, L. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the 8th Conference on Natural Language Learning (CoNLL-2004)*. Boston, USA, 89–97.

Carreras, X., Màrquez, L., and Chrupa, G. 2004. Hierarchical recognition of propositional arguments with perceptrons. In *Proceedings of the 8th Conference on Natural Language Learning (CoNLL-2004)*. Boston, USA.

Charniak, E. 2000. Maximum-entropy-inspired parser. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 132–139.

Hacioglu, K., Pradhan, S., Ward, W., Martin, J. H., and Jurafsky, D. 2004. Semantic role labeling by tagging syntactic chunks. In *Proceedings of the 8th Conference on Natural Language Learning (CoNLL-2004)*. Boston, USA.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics 19,* 2, 313–330.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM 38,* 11, 39–41.

O'Hara, T. and Wiebe, J. 2003. Preposition semantic classification via treebank and FrameNet. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003)*. Edmonton, Canada.

Punyakanok, V., Roth, D., Yih, W.-T., Zimak, D., and Tu, Y. 2004. Semantic role labeling via generalized inference over classifiers. In *Proceedings of the 8th Conference on Natural Language Learning (CoNLL-2004)*. Boston, USA.

Tjong Kim Sang, E. and Veenstra, J. 1999. Representing text chunks. In *Proceedings of EACL '99*. Bergen, Norway.